DOCUMENT RESUME

ED 146 539 CS 003 620

AUTHOR Calfee, Robert; Juel, Connie
TITLE How Theory and Research on Reading Assessment Can Serve Decision-Makers.
SPONS AGENCY Carnegie Foundation for the Advancement of Teaching, New York, N.Y.
PUB DATE 77
NOTE 35p.; Paper presented at the Minnesota Perspectives on Literacy Conference (Minneapolis, Minnesota, June 1977); See related document CS003621

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Decision Making; Diagnostic Tests; *Educational Assessment; Elementary Secondary Education; *Models; *Reading Processes; *Reading Research; *Reading Tests; Testing Problems
IDENTIFIERS *Interactive Reading Assessment System; *Minnesota Educational Assessment Program

ABSTRACT

After reviewing the information that teachers and other decision makers need to have about student achievement and some recent advances in the theory and practice of reading assessment, the author makes a number of recommendations for improving assessment programs. These include: do less massive, broad-band testing, but improve the reliability and informativeness of what testing is done; look to instruction as the model for what to test, and then consider the influence of the testing situation, the tester, and the materials; be sure the information will be organized in a useful way, around theoretical models of the reading process. Too often, decision makers have the option of too little information (a single test score) or too much information (a myriad of behavioral objective scores). In developing these arguments, the Minnesota Educational Assessment Program and the Interactive Reading Assessment System are analyzed to illustrate both problems and alternative approaches. (AA)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED146539

HOW THEORY AND RESEARCH ON READING ASSESSMENT CAN SERVE DECISION-MAKERS[1]

by

Robert Calfee

and

Connie Juel

Stanford University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Robert C. Calfee
Connie Juel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND THE ERIC SYSTEM CONTRACTORS."

[1]Preparation of this paper was supported in part by a grant from the Carnegie Foundation. We are grateful to Priscilla Drum, Dorothy Piontkowski, Barbara Tanner, Kay Thoresen, and Barbara Tingey for their assistance.

CS 003 620

HOW THEORY AND RESEARCH ON READING ASSESSMENT CAN SERVE DECISION-MAKERS[1]

Robert Calfee and Connie Juel, Stanford University

The controversy over educational testing continues to make headlines in newspapers and bold type on the covers of professional journals. The actual source of the unhappiness varies somewhat from one complainant to another, e.g., cultural bias, cost and time, need, etc. But frequently people express concerns with the inappropriateness of present measures of school achievement. Because of the perceived importance of reading to success in other school subjects, reading tests are challenged with special force. Parents are generally (not always) mystified by achievement tests but believe that their child's measured performance means something--it does. Teachers often express the feeling that present reading achievement tests don't measure what they teach--they are generally right. Administrators hope that their achievement scores will go up rather than down--about half the time their prayers are answered. (Hopes and prayers are the proper terms, for administrators are hard-pressed to find clear evidence that helps them act to improve reading scores.) Finally, school board members and legislators must feel frustrated that with so many resources being allocated to the improvement of reading, there is no clear trend toward improvement nationwide that matches the resources--and then they hear experts say that no one knows what the tests measure anyway!

What Are the Answers to all these Problems?

First - Existing reading tests do serve a useful purpose. As unidimensional indicators, they predict performance on other tests with remarkable accuracy.

Second - Present tests are extraordinarily inefficient--we could probably cut time and effort to one-fourth or even one-tenth

and obtain the same information. Group-administered, multiple-choice tests one-quarter as long, administered to half the number of students, would provide adequate data for the purposes for which they are suitable, at considerable reduction in cost, time, and effort for everyone.

Third - Present tests are <u>not</u> appropriate for all the uses to which they are put--they predict general success or failure in school, they call for action, but they do not tell what action to take.

We suspect that a lot of energy presently goes into teaching children what they already know (Durkin, 1974). Given limited resources, this ineffiency is troubling. Tests could provide evidence to highlight areas of need, but to do this requires tests that generate differential profiles--that reveal relative strengths and weaknesses--for student, for class, for school, and for district. Existing standardized achievement tests do not provide reliable profiles. Subtests exist, to be sure, but subscores with quite different labels are highly correlated with one another, and investigators have shown that the "profiles" are unreliable. Our own work has shown that existing methods of establishing total test reliability are inimical with the creation of tests that yield reliable profile information. Criterion-referenced and behavioral-objective tests have different kinds of problems. Aside from the fact that many look like standardized achievement tests, the teacher and others face a wide array of unorganized data (e.g., school administrators have found it difficult to make use of the findings of National Assessment of Educational Progress, U.S. GAO, 1976).

We are convinced that differential profiles exist, and that we have the methodology to measure them. Advances in differential diagnosis

for decision making are close at hand, in our opinion. These advances will build on two recent developments--adequate theoretical models of the reading process, and redefinition of the concept of test reliability.

Theoretical Models of Reading

So many diverse theories of reading exist, and they have proven to be of such little use in solving practical problems, that you might wonder why we turn at this juncture to the notion of theory. Nothing is so practical as a good theory, it is said, and in many areas of applied science this epigram has proven true. In principle, an adequate theory of reading should point us to appropriate methods of test design and construction, and should direct us to proper techniques of analysis and interpretation of the results. I believe that we can find such guidance in a theory, though probably not from the complicated models that many have proposed (Calfee, 1975).

The independent-process theory which we will describe below appears unduly simple, but it has powerful consequences (Calfee, 1976). And though lacking the intricacies of a computer simulation model or the elegance of a mathematical derivation, it does have practical consequences.

Independent-process theory rests on the assumptions that the mind carries out certain activities through the operation of independent cognitive processes--by analogy, the mind operates like a "works-in-a-drawer" television, rather than through a complexly interwoven and interactive network of processes (Figure 1). Psychologists and educators often say that people are complicated--and they probably are in some ways. But for certain purposes, including the design of reading assessment systems, a few simple categories of mental processes may suffice to describe the most important features of performance--that is the essence of the independent-process assumption.

- - - - - -
Figure 1
about here
- - - - - -

As will become apparent, we also think that the categories of independent processes in reading are closely linked to what is taught--in the case of derived skills like reading, people learn what they are taught, and they learn independently what they are taught independently. Thus, as applied to reading, the theory assumes the existence of separable skills, like decoding, vocabulary, and comprehension. To assess these as independent skills, we need clean subtests, which minimize the contribution of ancillary skills. We also need to introduce systematic variation in the content and context of testing, and the critical data include comparisons between performance in one set of circumstances and another.

An Example of an Independent-Process Model

Let's see how the principle of process-independence applies to the assessment of a student's ability to "read" and understand single words. The task we have in mind is a common one at the primary school level. The student is shown a list of words selected to represent a particular "level of difficulty." He is asked first to pronounce each word, and then to demonstrate that he understands a common meaning of the word.

What thought processes must the student bring to bear on the task in order to perform successfully? What are possible patterns of failure, and what do these patterns mean for instruction? The information-processing model for test design (Figure 2) incorporates three processes--attention, decoding, and lexical interpretation. We will look at each of these in turn.

- - - - - -
Figure 2
about here
- - - - - -

First, we consider how the student attends to the task. This process is a complex entity in its own right, including the overall level of activity, the extent to which the student selects relevant cues and rejects irrelevant information, and the degree to which the student can concentrate

the maximum available mental capacity on the task (Piontkowski & Calfee, in press). For our present purposes we will lump all of these into a single "box." We plan to influence this process by variation of a general character, and we will measure it in generic fashion. We include this process in the model because it seems likely to influence the operation of the other two processes, and because specialists in learning disability have identified attentional dysfunction as an important reason for reading failure (Ross, 1976). The design of the assessment system allows us to test this hypothesis for each individual student.

The second process, decoding, handles the translation of print into spoken language. Undoubtedly, there are subprocesses that handle specific aspects of the translation task, but for our purposes we again consider these as an aggregate.

The third process, lexical interpretation, refers to the student's ability to demonstrate a common meaning of a word presented in isolation. One may argue, and rightly so, that during the silent reading of connected prose, the student thinks in a quite different manner than when he is shown a word in isolation and asked what it means. The point is well taken, but irrelevant to the present situation. Students are asked to do both tasks as part of learning to read, and the high correlation between performance on the two tasks suggests that they share a number of elements in common.

Once the model is specified, the next step in test design is to designate one or more factors--variations in testing conditions--that are likely to strongly influence the operation of each process. An example of a relevant factor is shown above each of the processes in Figure 2. For instance, in the case of attention, it seems to us that the operation of that process should lead to better overall performance when the student

is individually tested in a quiet room than when he is tested with a group in a noisy room. We also propose that regularity of the letter-sound correspondences of the stimulus words should affect the decoding process, and that familiarity of the words should influence the lexical interpretation process. The design of the test includes all combinations of the factors, and so each student is tested under all combinations. Thus, in one set of situations the student is taken into a quiet room and asked to pronounce and to define words from combinations of letter-sound regularity and familiarity. The testing is then repeated with different words from the same design in a regular, noisy, crowded classroom.

Having specified variations that influence each process, we next want to find a way to measure the operation of each process. We recommend choice of the most direct measures possible. Thus, in addition to recording the correctness of the Pronunciation and Definition, the tester also records the student's Concentration on the task as a general measure of attention.

The purpose of the design variations is to measure the student's performance under different conditions, as a way of discovering relative strengths and weaknesses. This principle is akin to the clinical tester who, besides noting a person's overall intelligence test score, also considers the difference between the verbal and performance subtests.

Reliability of Profiles

Most tests are designed to optimize the reliability of the total score (Cronbach, 1976), The procedures we are proposing emphasize differences as much or more than overall summary scores (Calfee & Drum, in press).

In general, reliability refers to the degree to which a measurement is consistently reproducible. We can consider the consistency in performance

when a person is tested with one form of a test and then retested with a slightly varied form. Several things have changed. The exact form and content of the test have changed. The student has probably changed. He may have learned something, he may have forgotten something, he may have a headache now that he didn't have earlier. All these sources of variability tend to influence the reliability in test-retest situations (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Test developers tend to emphasize within-test reliability. There are several ways of thinking about this form of conistency (Cronbach, 1970, Ch. 6). For instance, suppose you divide the test items at random in two and correlate the two subscores. Repeat this operation for all possible split-half divisions of the test, then compute the average correlation between the half-scores (Cronbach, 1951). This provides a measure of the extent to which each item contributes consistently to the total test score. One way to obtain "perfect" intratest reliability is to use a test in which the items are so homogeneous that the student either fails or passes all items. Test developers, to the degree that they strive for high levels of intratest reliability, are under pressure to eliminate test items that yield divergent patterns of performance from one student to the next. The items that remain seem likely to measure general performance characteristics rather than performances that reflect specific instructional outcomes. So if you want a perfectly reliable test, ask the same question twenty times. Either a student knows the answer or he doesn't. This would be absurd, of course, but in the limit it is the "ideal" toward which reliability aims.

Maximizing intratest reliability is important when the test score is to serve for a major decision, but it may be counterproductive for instructional decision-making. Teachers need to know more than the student's

general ability. Individualization requires knowledge of diverse patterns of performance on specific tasks for different students. For the teacher, a "reliable" assessment instrument is more properly defined as one which accurately and consistently indicates the specific patterns of instruction that best fit the student's needs and capabilities.

We have discussed elsewhere detailed techniques for measuring the reliability of profiles, and have illustrated the application of these techniques to the design and analysis of reading tests (Calfee and Drum, in press). The technical details are not relevant to our present purposes, but several points deserve emphasis. First, differential information about strengths and weaknesses in separable skill areas is needed for intelligent decision-making. Second, in the design of most current reading tests, "the reliability of the test" is established in a way that optimizes item consistency with the total score. However, to obtain differential protile information requires the development of tests where profile reliabilities are optimized. Third, we suspect that increasing the reliability of patterns will require test developers to minimize generalized task demands and place emphasis on specific task demands in the construction of tests. Such steps should enhance the validity of the tests in significant ways.

Evaluation of the MEAP

In this section we apply some of the previous ideas in a critical evaluation of the Minnesota Educational Assessment Program (MEAP) (Minnesota Department of Education, 1974). The stated purpose of the Minnesota Assessment was to "examine the reading performance of Minnesota students, and determine which factors appear to account for a variation in that performance. This report, and analysis of the results, gives a clearer picture of how well students are reading and examines how groups of stu-

dents vary in performance. By describing the levels of reading performance in Minnesota, the report presents to educators, policy makers, and the lay public reliable information to use in the consideration of alternative directions for educational policy" (Minnesota Department of Education, 1974, p. 1).

The Minnesota Assessment is a generally fine piece of work of this genre. A variety of tasks and content are represented in this group administered, multiple choice test. The items are clearly laid out, and the instructions fairly readable. A detailed analysis of the results was carried out by several independent groups. At times the report has an air of "committee writing," but this is inherent in a multiple perspective approach. We did not have access to specific item analyses--if these are not available they would constitute an important addition to the report.

We will focus our critique on three points: (a) the relation between program goals and the content and analysis of the data, (b) the test layout and item construction characteristics, and (c) the formal of presenting measures and reporting data.

In a separate flyer (MSEP--Minnesota Statewide Educational Assessment Program), the following questions are raised:

> How many students in your school district or in the state can read fluently enough to be considered basically literate? How many cannot? How many students read well enough to deal with materials demanding critical, judgmental reading skills? How many students read well enough to be successful in a college setting? Are their ambitions in line with their abilities?

These are good questions, and sufficiently important to deserve validated answers. Unfortunately, no attempt is made anywhere in this large scale

data collection and analysis effort to validate the results. We will not spend long on this point--simply stated, it is crucial to obtain other kinds of information on success in schooling as a validating criterion for test instruments of this sort. All that we know from this assessment project is how well the students do on tests.

A second point about the program goals concerns the way that answers are provided. The test items from the Minnesota Assessment were used to generate several indices: Basic Literacy, School Success, Reading for Critical Evaluation and Citizenship, and Reading for Success in College. These indices are reported for aggregated data separately for each index. That is, one can find average performance in Basic Literacy as a function of various demographic and ethnographic factors. However, nowhere is information provided in a contrastive form, so as to show the relative strengths and weaknesses in these areas for various subgroups in the population. The reader can put some of the information together from the report to highlight these strengths and weaknesses, but the report doesn't do this job. It is not much of a secret to find out that low SES minority groups do poorly in virtaully all of these areas--what we also need to know is the character of their relative strengths and weaknesses. The report comes close to providing such information in Chapter 4, where "domain" averages are given for several categories of factors. Two samples of data for nine-year-old students are plotted in Figure 3, and it appears that the sharpest group differences show up in the comprehensive tasks. However, averages can actually obscure underlying patterns, and what is needed are actual profile statistics for students and schools (Calfee, 1976; Calfee & Drum, in press) Incidentally, the Minnesota Report is skimpy on descriptive statistics, like sample size, measures of variability and correlations, which could

- - - - - -
Figure 3
about here
- - - - - -

provide a more complete picture of the results). Nonetheless, the two profiles in Figure 3 suggest an interesting difference between the effects of variation in SES (a relatively sharp contrast in Passage Comprehension, compared with the other differences), and variation in Attention (fairly constant decrements for the Low Attention group in all domains). We think that information of this sort, sharpened and highlighted, could provide a more useful basis for action than separate compilations of test scores.

The reader may wonder why this is not a problem of "reporting." The answer is, it is a conceptual matter, and not simply a question of how to present data. Decision-makers at various levels need to begin thinking more about what students can and cannot do in particular instances, rather than focusing on overall levels of skill or weakness. For too many years, the student's "average" performance, weighted to favor verbal and academic skills, has served as a basis for making an overall judgment about that child. It is possible to highlight the child's strength (Cohen, 1973); it may be vital to deal with specific weaknesses. Thinking in this fashion is also likely to lead to tests that are designed for optimally reliable distinctions between significantly different areas of skill and knowledge.

Next, let us look at some of the items in the Minnesota Assessment. In Figure 4 is a set of eight items testing knowledge of prefixes and suffixes. We have several questions. First, why spend so much time testing the concepts of "prefix" and "suffix"? Surely, other questions about morphology are equally or more relevant to the child's level of vocabulary competence. Each item takes some time and energy--what other variations in content and task could be substituted to yield additional information about the students skill and knowledge? For instance, one might ask the student to add affixes that produce changes in meaning, or to show a knowl-

- - - - - -
Figure 4
about here
- - - - - -

edge of how an added affix changes meaning. Second, what does an error mean? The report includes some efforts to analyze error patterns, to be sure. Nonetheless, the character of the items makes it difficult to know precisely how to interpret an error. For instance, what if the student hasn't learned these two "reading jargon" terms, but knows the underlying concept of affixation? He is likely to miss all of the items, leading to the mistaken conclusion that he understands nothing about the concept.

Item content and task demands are important determinants of the concept of testing. If the student believes that the test requires him to look for "prefixes and suffixes" without regard to meaning, then the student does well to check anything that might be a prefix or suffix. For instance, mis and un are both prefixes sometimes, and the student might spot these in Items E and H. Nothing in the testing situation requires the child to check to see whether such a judgment makes semantic sense. A quick visual scan leads to errors that are promoted by the test design. It is easy to "design" items that promote errors--it takes considerably more planning and tryout to find the conditions that promote success.

It is hard to overemphasize the influence of testing context. For instance, in the report we read: "In the 'ignore the text' strategy, a student seemingly reads the question and chooses a distractor which represents common, but often inaccurate, knowledge" (p. 33). A thoughtful reader of the comprehension questions in the Minnesota Assessment might wonder why a student would follow any other strategy. Many of the questions hinge on external knowledge; more often than not, the student will be correct if he answers on the basis of external knowledge. Reading the prose wastes time, and adds little useful information. After enough instances of this sort, the clever (or lazy) student will conclude that he should look first at the questions, and only when uncertain return to the text.

- - - - - -
Figure 5
about here
- - - - - -

The exercise from the Minnesota Assessment in Figure 5 presumably is designed to tap vocabulary knowledge. However, the key to these questions is conventionality. For instance, one might believe that zebras are nervous all over, unlike horses. The student who is not familiar with real zebras might also think that they are relatively hairier than a horse. The student with some experience with zebras (picture books that stress the stripedness of this animal) will be at an advantage. Item B is even more dependent on conventionality (and sexism as well). We all know the mother's role includes sewing torn pants. A less conventional mother might decide to fold them up and put them aside--the problem is Billy's, not hers. The thoughtful and creative child might select "I don't know" as the best answer. But conventionality dictates that "I don't know" is never a proper answer on a test.

For each of these items, the critical question is, what is being tested? What does an error mean? What action should be taken by the decision-maker, teacher, or lay person when confronting a group of students (or individuals) who make mistakes on these items?

We will not follow this line further. However, for any test of this general character, we believe it is a good idea to ask continuously: What does the child have to know in order to succeed? What interpretation is to be put on a failure? How can the test item be modified to gain a wider range of information about the student's capabilities, and to ensure that the skill and knowledge being tapped is measured in as clean, precise, and uncontaminated fashion as possible?

The last point we want to make about the Minnesota Assessment concerns reporting data in a way that make them useful to decision-makers. The Minnesota report contains a great deal of information. Several efforts

have been made to simplify the presentation, and to reduce the tremendous amount of quantitative information. Decision-makers need descriptive information in the simplest possible form.[2]

Our complaint comes from the intrusion of unnecessary jargon and acronyms. For instance, in Figure 6 is a portion of a table from the report intended to show the relation between background factors and reading performance. The information is interesting and relevant, but translation into a comprehensible form is a time-consuming task for the expert, and probably outside the competence of many of the "educators, policy makers, and lay public" for whom the report is intended. Our point is simple--researchers who prepare reports should keep the audience in mind.

Perhaps some of our points may seem niggling. However, we are firm in the opinion that researchers and evaluators do have important information to convey to policy makers and the general public. Many are skeptical about the value of educational research and evaluation. This skepticism partly reflects the complexity of the phenomenon. We feel that it also reflects the failure of those who design, administer, analyze, and interpret test results to do their best to provide useful information in a clear manner. For better or worse, those of us who perform this task must be right in everything we do for our work to be of value.

What a Teacher Needs from a Test

In the preceding sections we have looked at characteristics of an assessment system that facilitate decision-making, illustrating the points by a situation where decision-making is at a fairly high level. Teachers also make decisions, and, in our opinion, the same principles apply at the level of the classroom as at the higher levels of state administrators and legislators. Partly because the individual classroom situation is more concrete and comprehensible, it may be easier to see the principles in action at that level.

What kind of information does the teacher need from a test if the goal is to improve instruction? First, the information is more useful if it points directly to the appropriate instructional treatment. Finding out that the student has not mastered the basic "long-short" vowel correspondences in English gives some direction to the teacher. Being told that the student "lacks adequate word attack skills" is less useful. And information that the student "cannot grasp the abstract character of letter-sound correspondences" may even be counterproductive--the teacher may try to teach "the abstract character . . ."

Second, test information should reveal the student's unique pattern of strengths and weaknesses, and not just his overall level of competence. The typical reading achievement test may inform the teacher that the student "reads at the 25th percentile," i.e., that seventy-five out of every hundred students in the nation do better on the test than this particular student. Or it may show that the student performs two grade-level-equivalents below expectation for his age. Such messages rarely surprise the competent teacher. If the student is, in general, doing poorly (or well, or average), the teacher does not need a standardized test to tell him so. Learning that an

old house you just bought is decrepit and in need of repair is no surprise--hopefully you knew that when you bought it. It is more useful to be told that the plumbing isn't as bad as it looks, whereas the apparently solid floor joists are riddled by termites and need immediate attention. Similarly, the teacher is helped by an assessment system that highlights patterns of relative strengths and weaknesses--such as a student's understanding of the meaning of certain words is relatively less well developed than his ability to decode them. Such patterns are often undetectable in performance on a generalized test, especially if the student performs poorly overall and the test is not appropriate to his level of competence (Calfee, Drum, & Arnold, in press).

Third, the teacher needs to be able to discover the conditions under which a student succeeds or fails on fairly specific tasks. A low score on a standardized test of reading achievement means the student has not given correct answers to many of the questions on a group-administered, multiple-choice test. To do well on such a test requires numerous skills; if the student fails, the test does not show which skills were lacking. For example, the usual comprehension task requires the student to have "gotten it all together"--it demands proficiency in word-attack skills, vocabulary knowledge, syntax, and ability to group the structural relations in the passage. Two students may both be labeled "poor comprehenders," but for different reasons. The label does not reveal the differences, and the teacher is left without the information needed to improve the situation.

The teacher can most easily determine the student's level of knowledge by asking him to perform the same basic task under a variety of conditions. For instance, perhaps the student who failed on the group administered comprehension test will succeed when the test is individually administered,

with care that the student understands the directions, that he reads the passage (and the questions and answers), and that he makes some response to every question. Or perhaps success comes only when the student is asked to read the passage aloud, and is given help on words he has trouble pronouncing. What if the student comprehends only when he is helped to understand words that, because of his level of language development, his ethnic background, or his particular interests and experiences, are unfamiliar to him? The student who comprehends when special care is taken to motivate him for the test doesn't need more instruction in the subject matter. His poor performance under regular conditions reflects something other than poor reading skill (Goodnow, 1972). Similarly, the student who can demonstrate understanding of a passage when he is helped to decode and define difficult words does not need further instruction on comprehension; he does need more training in decoding and vocabulary.

A fourth requirement for a test, if it is to be useful to the teacher, is that information is cheap and efficient. Administration, scoring, and interpretation must be quick and easy. Otherwise, the teacher is unlikely to use the test, even though it gives helpful and relevant information. The problem here is what Cronbach (1970, pp. 602ff) calls the bandwidth-fidelity dilemma. The need is for a test that covers a broad range of skills and knowledge, the provides variation in the task requirements, that has a "bottom" low enough and a "top" high enough for the variety of students and the extent of learning over the school year. Meeting all these criteria is not simple; however, we believe it is possible to design reading tests that come close to meeting these requirement.

A Practical Example--the Interactive Reading Assessment System

Our effort to apply these principles heuristically to improve reading test design, relying on our intuitions about underlying mental processes, is exemplified in IRAS (Interactive Reading Assessment System, Calfee & Calfee, 1977). Concepts of test design will be illustrated by the section of IRAS that measures comprehension skills.

We begin by laying out some of the major dimensions which influence performance on a comprehension task. To be sure, comprehension is a complex activity, involving numerous processes. Debate will continue for some time about what the term really means, how exactly to measure it, and what tasks to emphasize under this rubric. For our purposes, the questions have been resolved in a practical manner. The basic comprehension task entails asking a student to read a passage aloud, and then to respond to questions designed to tap his ability to extract specific details of information contained in the passage, to grasp relations among the facts, and to provide a reasonable summary of the main themes. We prefer to have the student read aloud, not because this is essential to comprehension, but because it provides direct evidence on the student's level of success in translating the printed text.

The first and most important dimension is the "difficulty" of the passage. If one collects a large sample of materials, appropriate to the interests and competence of elementary school children, these can be reliably graded by experts according to the relative ease with which students can read the passages. The features that enter into this dimension are partly known at present--among these are passage length, familiarity of the vocabulary (frequency of occurrence of words in print), syntactic complexity, number of propositions, and degree to which the passage deals with topics that arise in everyday experience, among others. Various readability formu-

las verify the existence of this dimension, and the degree to which one may reliably place a particular passage somewhere on the scale (Gilliland, 1972; Klare, 1974). These details are unimportant to our purposes, which are satisfied by the selection of a wide variety of passages that vary in difficulty, whatever it means.

A second important dimension for our purposes is the difference commonly referred to as "reading" versus "listening" comprehension. On the one hand, the tester can ask the student to read the passage himself and then test his understanding, or the tester can read the passage for the student, encouraging him to scan the material as it is read, and then can test the student's understanding. If the student fails when he reads for himself, he may still do well with similar materials when the tester reads for him. This contrast in performance has important implications for instruction, especially when compared with a third outcome where the student does poorly even when the material is read to him.

A third dimension, occasionally mentioned in test manuals but seldom part of either test validation or interpretation, is the character of the question asked. As noted earlier, one may ask the student to recall details of a specific proposition, to put together relations between propositions, or to summarize the structure of the passage--various other possibilities exist, but these are the main kinds of questions referred to in most discussions of how to measure comprehension (Guzak, 1972). Surely the tester/teacher would want to distinguish between the student whose ability to handle a comprehension task was weak for all categories of questions, and the student who regularly "got the facts" but couldn't organize them.

Another dimension closely related to the type of question is the response required--productive versus receptive. If the student is asked,

after reading a murder mystery, "Who committed the crime?" he must generate the answer on his own, reaching into memory for possible alternatives, then choosing the one that seems most plausible. If the student is asked, "Was it the butler or the grandson?" searching memory for viable alternatives is unnecessary. Only recognition is required, and the student may actually use his knowledge of the world to make the choice without reading the material at all--how often does a writer have someone murder his grandfather; it must have been the butler.

Figure 7 about here

Figure 7 shows how these dimensions are represented in sample materials from IRAS. For efficiency, the student is asked to help locate his level of competence. He looks at a graded series of passages (those in the figure are from the fifth and ninth levels in a series from 1 to 14), and tells the tester when he has reached a passage that he thinks he cannot read. The tester then asks the student to read the preceding passage aloud, and to answer several questions. If the student's reading performance is poor, or if he fails to answer the questions satisfactorily, the tester then asks him to read the next easier passage in the series. This procedure is continued until the student achieves a satisfactory level of performance. If the student is successful on the first passage, he is asked to read the next more difficult passage, and so on until he reaches a level which is too difficult for him.

The interactive procedure described above permits a rapid evaluation of the student's level of competence, and the degree to which performance changes with the difficulty of the text. After the reading test is completed, the tester then presents the student with a passage one difficulty level above his limit, and asks the student to follow along as the passage is read to him. Comprehension questions are asked in the usual fashion, and succes-

sively more difficult passages are presented until he fails to answer most of the questions correctly. The limits of listening comprehension are thereby established, which measures the contrast between reading and listening comprehension.

Examination of the questions in Figure 7 reveals a structure that includes variation in type of question, and productive versus recognition response demands. For questions 1 to 4, the sequence ranges from specific details through a summarization. The fifth question in each series places a different demand on the student--he must answer a question that is not answered by the passage, using knowledge that is assumed by the writer to be part of the reader's experience, and that is important for full understanding of the passage. On the one hand, it is reasonable to ask that most comprehension questions should be passage-dependent (i.e., should be based on information contained within the passage). But it is equally true that virtually anything a person reads makes sense only as external knowledge is brought to bear for interpretation (Bower, 1976). In IRAS, a sample of such knowledge is tested explicitly.

On the surface, IRAS resembles informal reading inventories and tests like the Gray Oral Reading Test (Gray, 1967). Indeed, portions of IRAS are modeled after procedures used in the Gray Oral. However, the design of the system permits measurement of contrastive difference scores, which can reflect relative strengths and weaknesses. Moreover, the incorporation of probes and explicit decision strategies serve to formalize the clinical features of the test. The tester not only is able to follow his nose, the test actually points the way. For instance, the tester is able to tell how the student performs when he is given a hint. For the student whose problem is lack of confidence rather than knowledge, it is important for the tester

to learn that he can do very well on a comprehension task when he is prodded, but fails when left on his own. This contrast in performance suggests that his problem has little to do with comprehension per se.

Summing Up and Some Recommendations

The theme in what we have proposed above is that a test ought to provide useful information about separable features of the collection of skills knows as reading. The level of detail in the breakdown of performance skills should depend on the decision-maker's need for information. To be useful, the information needs to have structure and organization. The district superintendent is not helped by being told the average percentage of correct responses for each of 394 behavioral objectives for each school in the district. For that matter, neither is the teacher likely to be helped by knowing the same information about each student in the class. To be told that the competence in decoding and vocabulary skills is relatively higher than competence in literal comprehension skills provides a more reasonable basis for action.

This theme may not seem to suggest much change from present procedures. After all, most standardized achievement tests provide a breakdown into subtests, do they not? There are two differences between what is being suggested in this paper and existing practices. First, present tests, shackled by the restrictions to group administration and multiple-choice format, do not provide adequate coverage of all the relevant areas of competence--for instance, we currently know little about what students know about decoding, because such information requires that the student read words or text aloud. Second, we have emphasized the influence of the testing context on performance, and the related matter of measuring corollary aspects of performance. If a student succeeds in performing a task under

one condition but fails under another, then basic knowledge is assured--it is the student's ability to apply that knowledge that is in question. Some students do poorly on a group test because of distraction and lack of motivation. Tested individually, with care to assure that they understand what is required, and with the attention and interest of another human being to motivate them, the same students may do quite well. From the point of view of someone who has to decide what action to take to help the student improve his performance, this latter information would seem quite important. Even in a closely monitored individual testing situation, the way that a student behaves may be an important piece of information. The student who is obviously concentrating, who tries alternatives when he suspects he is wrong, whose posture and wrinkled brow show dedication to the task--and who still fails to perform well--requires different treatment from the student who fails and who also exhibits obvious lack of attention, hyperactive movement, or disinterest.

There is only a modest amount of research directly based on the ideas in this paper, and so recommendations for action should be received with caution. However, based on our knowledge and experience in reading research, we feel relatively confident in presenting three concrete recommendations that depart substantially from present practices:

--Do less massive, "broad-band" testing, but improve the quality (the reliability and informativeness) of what testing is done (Venezky, 1974). Don't do away with all testing, for that weakens accountability.

--Look to instruction as the model for what to test, and then consider the influence of the testing situation, the tester, and the materials. For the teacher, the best question is often, "Are there any

conditions under which the student can succeed at this task?"

Reading teachers teach several different things, and what is taught will vary from one level to another. Assessment systems should be designed to reflect these differences, and the emphasis should be on the reliability of the patterns of these differences.

--Information must be organized if it is to be useful. Too often, educational decision-makers have the option of too little information (a single test score) or too much information (a myriad of behavioral objective scores). Theory provides a useful tool for organizing knowledge. In reading we know enough about the phenomenon to build models of the process that are of practical value in creating tests and interpreting test data.

These recommendations build on the assumption that if viewed properly, the acquisition of reading follows a small number of fairly simple themes, and that assessment reflecting these themes in a straightforward fashion can serve directly for decision-making. Research on reading abounds. Much of it portrays reading as a complex of interactive skills, idiosyncratic to the individual student-teacher-school combination. Such a description may be partly true--it certainly lends itself well to the creation of intricate flow charts and complex computer programs. But we think that reading is perhaps not no intricate after all. Teaching a child to read is sometimes a demanding task, but many teachers succeed at this task year after year--success in this endeavor is certainly more common than success in teaching a computer to read.

The presumption of "complexity" goes against the canon of parsimony, but more troubling, it leaves us unable to take action--experience is a poor guide when every situation is unique. The concept of independent

processes is simple and practical, and readily serves as a basis for action. Research is paying off. There have been some false leads, and progress has seemed slow at times. But we believe that the next ten years will see some significant breakthroughs in the assessment of reading--we are seeing some useful results already (e.g., McDonald & Elias, 1975). We are not about to solve all of our problems--curriculum development and teacher training will not be immediately influenced by improved assessment techniques. But the availability of a richer information base from the differential assessment of reading skills will leave decision-makers at all levels in a better position to find out where they need to take action.

Footnotes

[1]Preparation of this paper was supported in part by a grant from the Carnegie Foundation. We are grateful to Priscilla Drum, Dorothy Piontkowski, Barbara Tanner, Kay Thoresen, and Barbara Tingey for their assistance

[2]The justification for some of the simplification may be questionable, to be sure--for instance, when one looks at the distribution of scores on the basic literacy index, it is not clear why making twelve correct responses to the eighteen questions should be identified as success, whereas making eleven or fewer correct should be failure. There are procedures for validating such decisions (Calfee, 1977), but these were not in force here. Moreover, the several indices and test batteries in the Minnesota Assessment may, in fact, be measuring a single underlying trait ("There were 12 measures of a school's reading performance level. Correlation analysis showed these 12 to be highly intercorrelated" p. 144) to a degree that the test is unidimensional, the analysis could need a single index, rather than the several provided. In fact, we suspect that the correlation is due to the fact that many of the items are not particularly clean, and that several of the indices were constructed to use overlapping items.

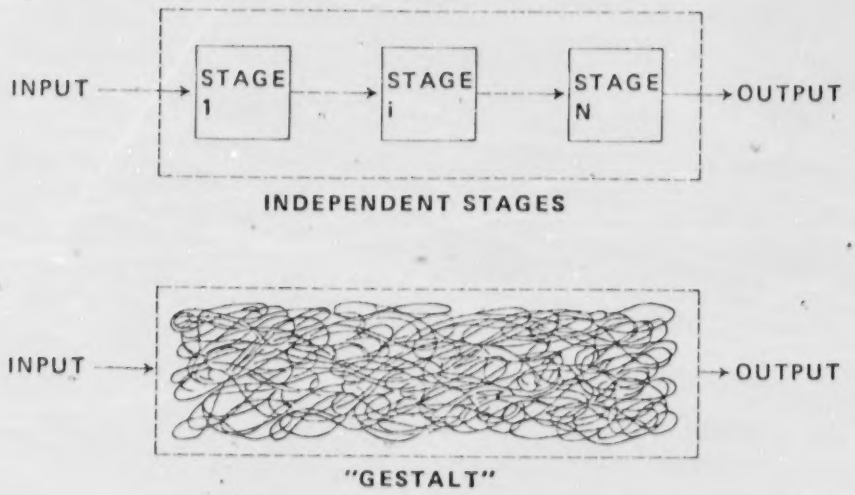

Figure 1. Independent-stage model of thought and contrasting Gestalt Model. Sequence of specifiable components operate on information according to independent-stage model, while intricate interactions are postulated in Gestalt model. After Calfee and Floyd, 1972.

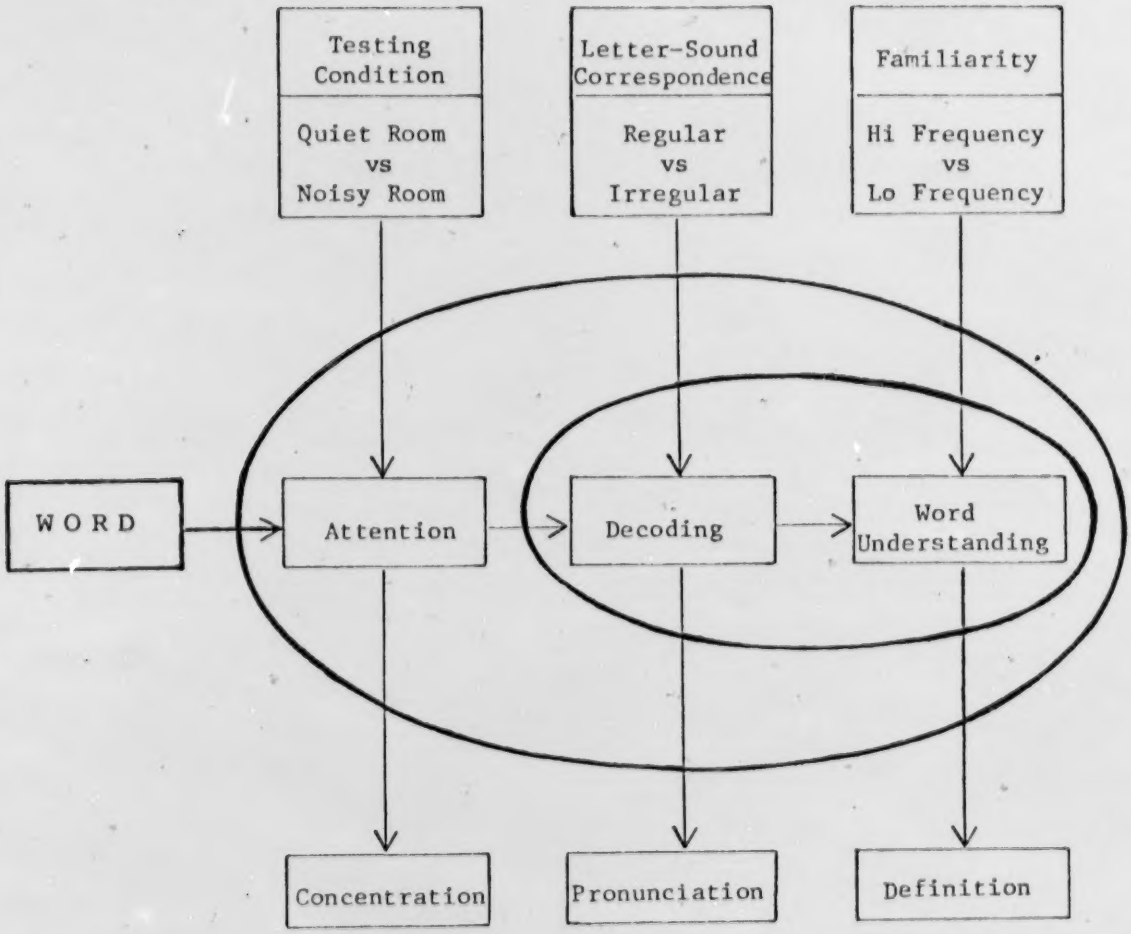

Figure 2. Information processing model for word knowledge task

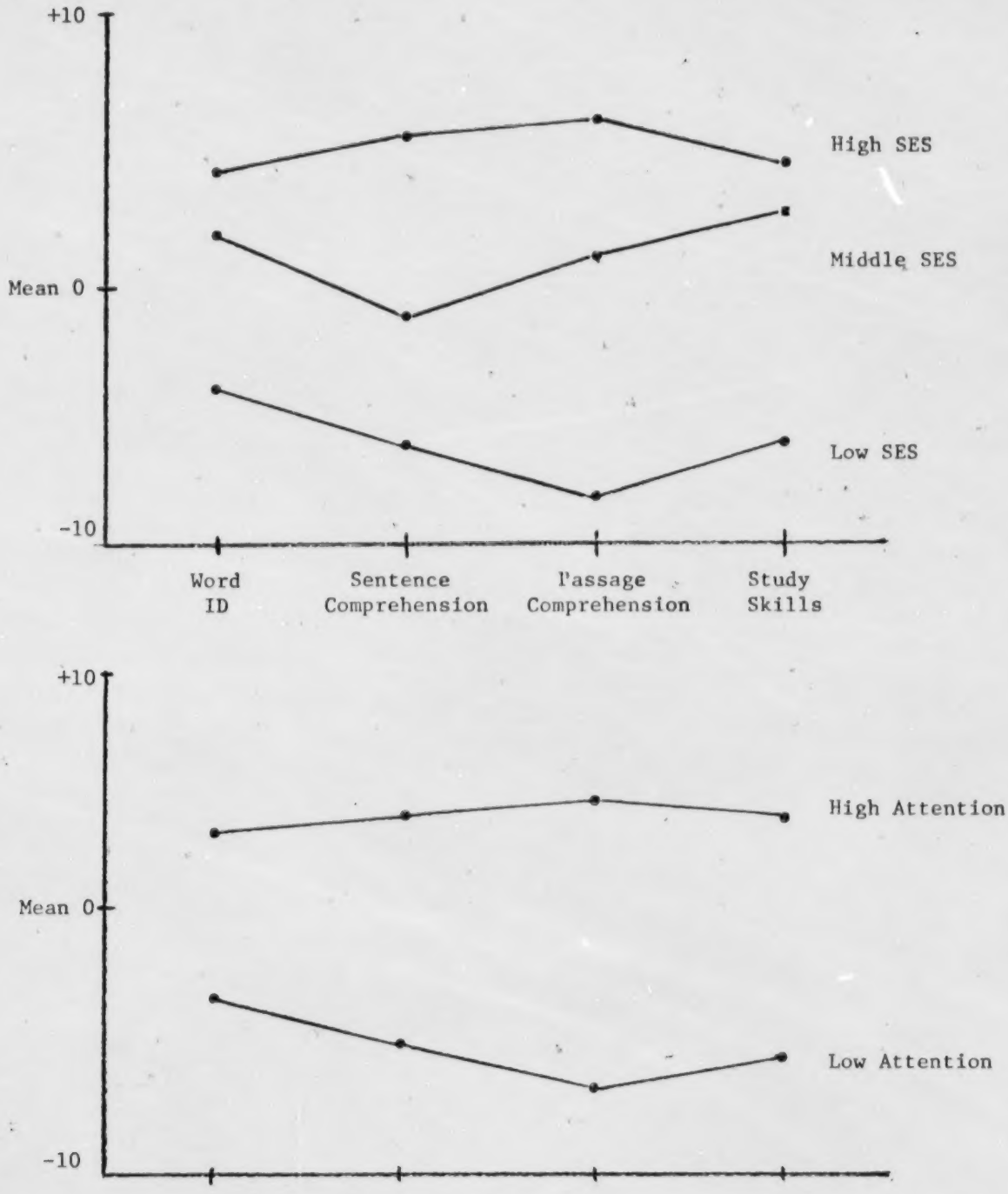

Figure 3. Patterns of group affects in 9-year-olds

In this exercise we want to see how well you can recognize a prefix or a suffix in a word. For each part read the key word and decide whether it has only a prefix only a suffix, both a prefix and a suffix, or neither a prefix nor a suffix. Then fill in the oval next to your choice.

Example 1	Example 2
The word run has	The word react has
○ only a prefix	○ only a prefix
○ only a suffix	○ only a suffix
○ both a prefix and a suffix	○ both a prefix and a suffix
● neither a prefix nor a suffix	○ neither a prefix nor a suffix
○ I don't know	○ I don't know

A. The word careless has

○ only a prefix
○ only a suffix
○ both a prefix and a suffix
○ neither a prefix nor a suffix
○ I don't know

B. The word disagreeable has

○ only a prefix
○ only a suffix
○ both a prefix and a suffix
○ neither a prefix nor a suffix
○ I don't know

C. The word discolor has

○ only a prefix
○ only a suffix
○ both a prefix and a suffix
○ neither a prefix nor a suffix
○ I don't know

D. The word impossible has

○ only a prefix
○ only a suffix
○ both a prefix and a suffix
○ neither a prefix nor a suffix
○ I don't know

E. The word mister has

○ only a prefix
○ only a suffix
○ both a prefix and a suffix
○ neither a prefix nor a suffix
○ I don't know

F. The word preheat has

○ only a prefix
○ only a suffix
○ both a prefix and a suffix
○ neither a prefix nor a suffix
○ I don't know

G. The word reddish has

○ only a prefix
○ only a suffix
○ both a prefix and a suffix
○ neither a prefix nor a suffix
○ I don't know

H. The word union has

○ only a prefix
○ only a suffix
○ both a prefix and a suffix
○ neither a prefix nor a suffix
○ I don't know

Figure 4. Exercise testing prefixes and suffixes. Data of Minnesota Educational Assessment Program (MEAP) 1974.

In this exercise we want to see how well you can use the clues given in a passage to select the best word to complete a sentence. In each part you are to read the passage and the four choices which follow it. Then decide which word best completes the sentence with the blank space and fill in the oval next to your choice.

Example 1

The sun had set and now everything was _____outside.

- ○ light
- ● dark
- ○ wet
- ○ warmer
- ○ I don't know

Example 2

They watched the dog _____ his fleas.

- ○ scratch
- ○ cat
- ○ feed
- ○ pet
- ○ I don't know

A. The zebra, unlike a horse, is _____ all over

- ○ striped
- ○ black
- ○ hairy
- ○ nervous
- ○ I don't know

B. After Billy tore his pants,he carried them to his mother and she _____ them up.

- ○ cut
- ○ folded
- ○ pushed
- ○ sewed
- ○ I don't know

Figure 5. Exercise designed to tap vocabulary knowledge. Data of Minnesota Educational Assessment Program (MEAP) 1974.

Variable	Level	No. Schools Sample 3	Means Y3
SPSESGE3	≤20%	68	59.4
	20-40	106	62.9
	>40	52	66.8
SREDMAT	≤50%	44	66.6
	50-75	129	63.6
	>75	53	57.3
SSCHLIB	≤30%	42	58.4
	30-60	142	63.4
	>60	42	64.6

	School Measures	Average Percent Correct Answers on Reading Test	No. of Schools
Socio-economic status	Percent students in school from high SES homes:		
	less than 20%	59	(68)
	20-40%	63	(106)
	more than 40%	67	(52)
	Percent students in school with limited reading materials in home:		
	less than 50%	67	(44)
	50-75%	64	(129)
	more than 75%	57	(53)
	Percent students using school library at least once a week:		
	less than 30%	58	(42)
	30-60%	63	(142)
	more than 60%	65	(42)

Figure 6. Illustration of reporting format from MEAP (Chapter 8) and easier-to-read format.

Note: In the original table, three samples are presented, but without any clear indication of how they differ. We present the data for Sample 3 only.

Dr. Albert Einstein's neighbor was worried. Every day her small daughter went to call on the great scientist. At last the mother went to Einstein. She told him she was sorry if the girl was keeping him from his work.

"Oh, not at all," Einstein told her. "I like her to come to see me. We get along quite well."

"But what could you and an eight-year-old girl have in common?" asked the mother.

"A great deal," said the scientist. "I love the jelly beans she brings me. And she loves the way I do her arithmetic lesson."

1. How old was the girl in the story? _____

   6, 8, or 9 years old

2. What two things did the girl bring to Einstein each day? _____

   Her violin and a letter from her mother; gum drops and her music book; her arithmetic lesson and jelly beans

3. What do you believe Einstein though about the lessons? _____

   Were they new, easy, or strange for him?

We talk about a dog being man's best friend, but as often as not it's really the other way around. My Great Dane, Max, for example, seems to thing of me as his pet.

To begin with, he is bigger than I am. Max stands seven and a half feet on his hind legs and weighs 280 pounds. There is something about a dog as big as a Shetland pony that keeps you from ordering him around quite as you would, say a poodle. But Max has gotten the idea that he was really meant to be a lap dog. He will come when I am asleep and lie across my legs, which makes it quite impossible for me to move until he want me to. And if he decides to sleep in a spot where I will stumble over him constantly--well, there is no moving him, of course.

1. How much does Max weigh? _____

   100, 200, or 300 pounds?

2. In this story Max's owner compares his size to another animal. Name the animal.

   A teddy bear, a poodle, or a Shetland pony?

3. How does the author feel about the saying "a dog is man's best friend"? _____

   He agrees with it; he thinks the opposite is true; he thinks it isn't true for Max

Figure 7. Examples of materials from IRAS for testing comprehension of narrative passages (portions of passages omitted). After Calfee & Calfee, 1977.